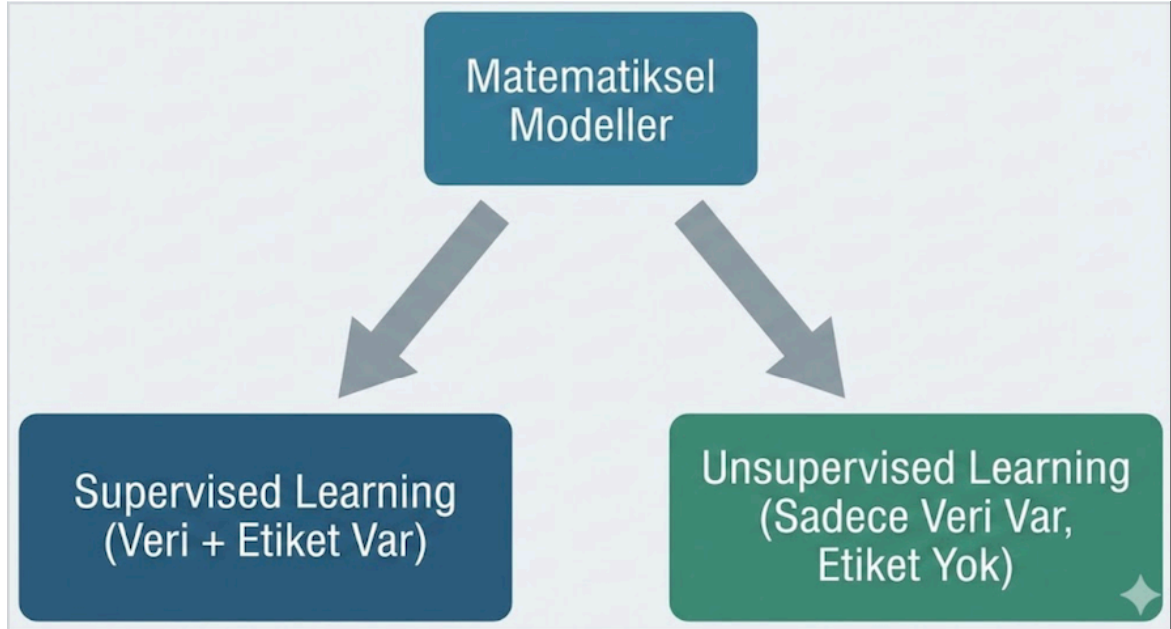


Stratejik Pivot: Neden Regresyon Değil, Sınıflandırma Modeline Geçiyoruz?

Projenin başlangıç aşamasında kurguladığımız, şirketlerin bağış potansiyel nokta atışı tahmin etmeye (Floating Point Regression) odaklanan iş akışını, veri gerçekleri ve modelin uygulanabilirliği nedeniyle temelden değiştirmemiz gerekmiştir. İlk kurgumuz olan "Linear Regression" tabanlı yaklaşımlar, ancak elimizde kesin sayısallaştırılmış (labeled) bağış verisi olduğunda çalışabilirdi. Elimizde türkiyedeki hiçbir şirket için bir "potansiyel sayı (atıyorum 100 üzerinden 80)" gibi bir label olmadığı için linear regression yapamayız. Çünkü linear regression supervised bir task ve "y eksenine" ihtiyacımız var.

Bu da bizi herhangi bir matematiksel model düşünüyorsak dünyaya açılmaya itti. Yani "Türkiyeden veri elde edemiyorsak neden dünya çapında verileri elde edip, standart hâline getirip sonrasında modelimizi türkiye çapında uygulamayalım?" Bu araştırma beni dünya çapındaki kanser şirketleri corporate partnerleri hakkında nasıl veriler paylaşıyor araştırmasına itti, çünkü ancak ve ancak elimizdeki verilere göre bir matematiksel model belirleyebilirdik. Öncelikle matematiksel modelleri 2'ye "Supervised ve Unsupervised Modeller" olarak ikiye ayırdık.



- Fakat unsupervised learning'de verilen kararları açıklamanın bir yolu yoktu, dolayısıyla model tarafından hangi kararın neden verildiğinin açıklamasını yapamayacağımız için sıkıntıya girecektik. Aynı zamanda supervised learning kısmında da neural network geliştirecek kadar verimiz yok, olsa dahi bunu seçmek mantıklı değildi çünkü neural networklerin de yine aldığı kararlar bir "blackbox" olarak bilinir ve açıklanabilirliği düşüktür.
- TKD için **Explainability (Açıklanabilirlik)** en az **Accuracy (Doğruluk)**

kadar önemlidir. TKD yönetimine gidip "Yapay zeka buna %90 potansiyel dedi ama nedenini bilmiyoruz" demektense, "Bu şirket perakende sektöründe, cirosu şu aralıkta ve daha önce sağlık STK'larına bağış yapmış, o yüzden skor yüksek" diyebileceğimiz bir model çok daha değerlidir.

Dolayısıyla supervised learning'i kesinleştirmiş olduk ve supervised learningde kesinlikle neural networkler yasaklanmış oldu. Şimdiki yeni soru ise bir multi-classification model mi geliştirmeliyiz yoksa floating point tahmin edeceğimiz bir regression mı yapmalıyız oldu. Burada ise dünya çapındaki kanser derneklerinin nasıl bir veri yayınladıklarına dependent bir karar vermek zorundayız.

Dünya genelindeki kanser derneklerini (ACS, CCS, St. Jude) incelediğimizde, verilerin sayısal bir süreklilik (continuous) değil, **kategorik bir hiyerarşi** (Visionary, Champion vb.) yayımlandığını gördük. Bu durum, projenin doğasını bir sayı tahminleme işinden, bir **Sınıflandırma (Classification)** problemine dönüştürdü. Ayrıca, elimizdeki veri setinin boyutu sebebiyle classification tasklerinde daha başarılı modelleri seçme yolunda az veri ile iyi çalışan modellere yönelmemiz gerekiyordu. Bu da bizi **Random Forest** veya **XGBoost** gibi ağaç tabanlı (Tree-based) modellere redirect etti. Böylece:

- Dünya çapında yayımlanan categorical verilere uyum sağlandı.
- Açıklanabilirlik konusunda TKD'ye cevap verilebilir hâle gelindi.
- Az veriyle güçlü modeller geliştirme konusunda sorun yaşanmadı, kaliteden ödün verilmedi.

Seçtiğimiz 2 potansiyel model anlaşıldıysa, şimdi de sürecin nasıl işleyeceğine bakalım. Bu sayede takım içi görev dağılımı konusunda da herkes genel bir fikre sahip olacaktır. Bu süreci yönetmek için **"Tersine Mühendislik ile Proxy Labeling" (Vekil Etiketleme)** stratejisini izleyeceğiz. İşte adım adım stratejimiz:

Adım 1: "Ground Truth" (Gerçek Veri) Nedir? (Data Reconnaissance)

Öncelikle, modelimizin neyi tahmin etmeye çalışacağını (Label/Y) belirlememiz lazım. Dünyadaki STK'lar genellikle **tam bağış miktarını (floating point)** yayınlamazlar (KVKK ve ticari gizlilik gereği). Ancak, raporlarda **Tier (Kademe)** sistemi yayınlarlar.

- **Veri Kaynağı:** American Cancer Society (ACS), St. Jude, Canada Cancer Society devlerinin web sayfaları.
- **Bulacağımız Veri Tipi:**
 - *Visionary Partners*: \$1M+ (Label: 5 - Çok Yüksek)
 - *Champion Partners*: \$500k - \$999k (Label: 4 - Yüksek)
 - *Leader Partners*: \$100k - \$499k (Label: 3 - Orta)
 - *Supporter*: \$50k - \$99k (Label: 2 - Düşük)
- **Zaten tam da bu nedenden ötürü, yani bir floating point değil de**

class yayımlandıklarından dolayı, bir classification task'ine dönüştürdük modeli.

Adım 2: Model Mimarisi Seçimi

Burada "**Random Forest Classifier**" veya "**XGBoost**" en mantıklı seçimdir. Neden?

1. **Küçük Veriyle Çalışır:** 200-300 şirketin verisiyle bile anlamlı sonuçlar üretebilir.
2. **Feature Importance:** Model sana şunu söyler: "*Ciro, çalışan sayısından %20 daha önemli.*" Bu, TKD için stratejik bir içgörüdür.
3. **Eksik Veri Toleransı:** Bir şirketin cirosunu bulamazsanız bile model diğer feature'lardan tahmin yapabilir.

Adım 3: Feature Engineering (Girdiler Ne Olmalı?)

A. Finansal Kapasite (Capacity)

- *Tahmini Yıllık Ciro (Revenue):* (LinkedIn/Glassdoor verisiyle scale edilebilir)
- *Çalışan Sayısı:* (Büyük şirketlerin başlı bütçesi genelde daha büyüktür)
- *Halka Açıklık Durumu:* (Halka açık şirketlerin şeffaflık ve itibar baskısı daha yüksektir)

B. Sektörel Uygunluk (Affinity)

- *Sektör:* (Perakende ve FMCG > B2B Sanayi. Çünkü B2C'nin halkla ilişkiler ihtiyacı daha yüksektir.)
- *Ürün Tipi:* (Fiziksel ürün satanların "Cause Marketing" yapma ihtimali daha yüksektir.)

C. Geçmiş Davranış (Propensity)

- *ESG Skoru:* (Varsa, yoksa web sitesinde "Sürdürülebilirlik Raporu" var mı? 1/0)
- *Daha Önce STK ile Çalışmış mı?:* (Web sitesinde "CSR" veya "Social Responsibility" sayfası var mı?)

Note: Aşağıda direkt olarak columnların ne olacağı verilmiştir.

Adım 4: Veri Toplama ve Model Eğitimi (Workflow)

Bu taski başlatmak için şu **hibrid süreci** izleyebiliriz:

1. Global Data Scraping:

- ACS, St. Jude ve Canadian Cancer Society'nin web sitesinden partner isimlerini ve bulundukları "Tier"ları çekelim. (Örn: Walmart -> Tier 5, Delta Airlines -> Tier 4).
- Bu bizim **Training Set**imiz olur (tahmini 200-300 şirket).
 - ♦ Bu projemizin ilk aşaması olmalı, dolayısıyla ilk görev bölümünü burada gerçekleştirebiliriz. Kendi aramızda 3 dev kanser derneğini bölüştürüp, her birimiz (şirket, label) tuple'ını oluşturulmalı ve sonrasında herkesin farklı derneklerdeki topladığı veriler bir araya getirilmeli. Böylece şirket ismi ve hangi büyüklükte başlı yaptığının class label'ı bir araya gelmiş

olur.

2. Feature Zenginleştirme (Enrichment):

- Bu şirketlerin (Walmart, Delta vb.) sektörü, çalışan sayısı ve ciro aralığını (Revenue Range) LinkedIn veya halka açık verilerden çekip tabloya ekleyelim.
 - ◆ Şu an şirketimiz ve labelları var, fakat hangi verilere göre modelimizin eğitileceğini belirlememize rağmen bu veriler elimizde yok. Yani yüzdelik finansal büyüklük, halka açıklık, ESG skoru, Çalışan sayısı gibi verileri sayfadan çektiğimiz her şirket için detaylandırmamız ve featureları doldurmamız gerekiyor. Bu da ikinci aşama olmalı task bölünmesi sırasında.
 - ◆ Burada da şirketleri ortak bir excel'e taşıdıktan sonra şirket bazında 4 kişi arasında 200'e yakın şirket bölüşülmeli, her birey 50'ye yakın şirket ele almalı. Her birey bu 50 şirketin çeşitli metriklerini internette araştırıp bulup excel tablosunu doldurmalı. Bu görevin sonunda ise verisetimiz tam hâline geliyor.

3. Model Eğitimi:

- Model şunu öğrenecek: *"Perakende sektöründe olup, 10.000+ çalışanı olan şirketler genelde Tier 5 (Çok Yüksek Potansiyel) oluyor."*
 - ◆ Verimiz tam olduğu için artık featurelar ile alakalı insightları öğrenmek modele kalmış, random forest ve XGBoost bunu yapacak. Biz featureları vereceğiz, o thresholdları akıllı bir şekilde öğreniyor olacak.

4. Türkiye Uygulaması (Prediction/Inference):

- Modelimizin eğitimini yaparken region-agnostik olmasına özen göstereceğiz. Dolayısıyla latin amerikada çalışan bir modelin türkiyede de uygulanabilir olması için özellikle finansal verileri yüzdelik olarak modele veriyor olacağız. Ki bu şekilde döviz farklılıklarını ortadan kaldıralım.
- Sonrasında inference/tahmin için ise hangi şirketin class'ını öğrenmek istiyorsak model geliştirildikten sonra o şirketin verilerini modelimize vereceğiz. Dolayısıyla modelimizin bir çıktı üretebilmesi için yine bu 9 tane aşağıda belirteceğimiz column'a ihtiyacı olacak. Örneğin Migros'un bağış potansiyelinin hangi segment'e girdiğini merak ediyoruz, migros hakkında bu "Feature Zenginleştirme" kısmında yaptığımız gibi TKD verileri toplayacak (zaten public), sonrasında bunu input olarak modele verip output olarak da bir class elde edecek.
- Model bakacak: "Migros, perakende, çalışan sayısı yüksek, B2C..."
-> **Tahmin: Yüksek Potansiyel (class y = 5).**

Partnerlerini 5 farklı şekilde classify eden dünyanın en büyük kanser dernekleri (200-300 firmalık veri sağlar)

CCS: <https://cancer.ca/en/get-involved/partnerships/our-corporate-partners>

- 5 - Çok yüksek: Visionary Partner
- 4 - Yüksek: Groundbreaking + Excellence Partner
- 3 - Orta: Inspiration Partner
- 2 - (Düşük-orta): Champion Partner
- 1 - Giriş: Guiding Partner + Caring Partner

STJude: <https://www.stjude.org/get-involved/other-ways/partner-with-st-jude/corporate-partners/companies.html>

- 5 - Çok yüksek: Vision Partners
- 4 - Yüksek: Hope Partners
- 3 - Orta: Dream Partners
- 2 - (Düşük-orta): Inspire Partners
- 1 - Giriş: Believe Partners

ACS: <https://www.cancer.org/about-us/our-partners.html>

- 5 - Çok yüksek: Visionary Partners
- 4 - Yüksek: Groundbreaker Partners
- 3 - Orta: Pioneer Partners
- 2 - (Düşük-orta): Champion Partners
- 1 - Giriş: Guardian Partners

9th Cycle - TKD Project: Corporate Donation Prediction Model Schema

Input Features (The "X" Variables)

Column 1: revenue_global_rank_percentile

- **Reason to include:** Dolar/TL kur farkından etkilenmemek için mutlak para birimi yerine, şirketin kendi ülkesindeki ekonomik sıralamasını kullanırız. "Zenginlik" kavramını yerelleştirir.
- **Data Type:** Float (0.00 - 1.00 arası. Örn: 0.99 = En tepedeki %1)
- **Importance: Critical.** Şirketin finansal kapasitesinin (Wallet Size) birincil göstergesidir.

Column 2: employee_count

- **Reason to include:** Ciro verisinin gizli olduğu özel şirketlerde (Private Companies) operasyonel büyüklüğü anlamının en güvenilir yoludur. İnsan kaynağı gücünü temsil eder.
- **Data Type:** Integer (Örn: 1500, 45000)
- **Importance: High.** Ciro ile koreledir ancak veri eksikliğinde "backup" görevi görür.

Column 3: is_publicly_traded

- **Reason to include:** Halka açık şirketlerin şeffaflık zorunluluğu, yatırımcı ilişkileri baskısı ve daha katı ESG hedefleri vardır. Bu da onları daha disiplinli bağışçılar yapar.
- **Data Type:** Binary (0 = Private, 1 = Public)
- **Importance: Medium-High.** Kurumsal ciddiyeti ve nakit akışı erişimini gösterir.

Column 4: years_active

- **Reason to include:** Şirketin köklülüğünü ölçer. 50 yıllık bir holdingin "Legacy Partner" olma ihtimali, 3 yıllık bir girişimin ihtimalinden farklıdır. Kurumsal olgunluğu (Maturity) temsil eder.
- **Data Type:** Integer (Örn: 45)
- **Importance: Medium.** Sürdürülebilirlik potansiyelini ayırt eder.

Column 5: industry_simplified

- **Reason to include:** Sektör, bağış motivasyonunu belirler. Perakende (Retail) "Cause Marketing" yaparken, ilaç (Pharma) veya Enerji sektörü itibar/kriz yönetimi için bağış yapar.
- **Data Type:** Categorical / One-Hot (Values: Retail_FMCG, Finance, Pharma_Health, Energy_Mining, Tech_Telco, Other)
- **Importance: High.** Şirketin "Neden" bağış yapacağını belirler.

Column 6: business_model

- **Reason to include:** Müşteri kitlesini ayırır. B2C (Halka satış yapan) firmalar marka görünürlüğü ve reklam için bağış yapmaya çok daha isteklidir.
- **Data Type:** Binary (0 = B2B, 1 = B2C)
- **Importance: High.** Pazarlama bütçesinden pay alma ihtimalini ölçer.

Column 7: has_esg_content

- **Reason to include:** Şirketin web sitesinde "Sustainability" veya "CSR" sayfası olması, bağış yapmaya yönelik "Niyetini" (Intent) kanıtlar. Parası olup niyeti olmayanı ayıklar.
- **Data Type:** Binary (0 = No, 1 = Yes)
- **Importance: High.** Şirketin kültürel uygunluğunu (Affinity) ölçer.

Column 8: linkedin_follower_count

- **Reason to include:** Şirketin dijital erişim gücünü (Reach) ölçer. Takipçisi yüksek firmalar, TKD için sadece finansal değil, aynı zamanda farkındalık ortağıdır.
- **Data Type:** Integer
- **Importance: Medium.** Cause-marketing kampanyalarındaki "reklam değerini" temsil eder.

Column 9: hq_region_type

- **Reason to include:** Karar vericinin nerede olduğunu anlar. Global merkezler (HQ) genellikle yerel şubelerden (Local Branch) daha büyük bütçeye sahiptir.
- **Data Type:** Categorical (0 = Local Branch, 1 = Global/Regional HQ)
- **Importance: Low-Medium.** Bütçe yetkisini tahmin eder.

Target Label (The "Y" Variable)

Modelin tahmin edeceği sınıflandırma (Classification) etiketleri:

- **5 (Very High Potential - Visionary):**
 - *Tanım:* Oyun değiştirici, ulusal çapta stratejik ortak. (Örn: \$1M+ veya Ulusal Ana Sponsor).
 - *Global Örnekler:* Walmart (ACS), CIBC (CCS), FedEx (St. Jude).
- **4 (High Potential - Strategic):**
 - *Tanım:* Yüksek bütçeli, yıllık taahhüt veren kurumsal ortak.
 - *Global Örnekler:* Delta Airlines, Wheaton Precious Metals.
- **3 (Medium Potential - Mid-Market):**
 - *Tanım:* Etkinlik sponsoru veya orta ölçekli kampanya ortağı.
 - *Global Örnekler:* Yerel bankalar, zincir restoranlar.
- **2 (Low-Medium Potential - Growth):**
 - *Tanım:* Potansiyeli olan ancak henüz hacimli bağış yapmayan, proje bazlı destekçiler veya ürün pazar yeri (Marketplace) katılımcıları.
- **1 (Entry Level - Mass Market):**
 - *Tanım:* Giriş seviyesi, tek seferlik bağış yapan veya sembolik destek veren KOBİ'ler.

Note on Feature Selection: Seçilen bu 9 özellik (feature), birbirine dik (orthogonal, independent) olacak şekilde tasarlanmıştır; yani her biri şirketin farklı bir boyutunu (Büyüklük, Niyet, Kültür, Erişim) açıklar. Örneğin, *Employee Count* ve *Revenue Rank* birbiriyle ilişkili görünse de, ciro bilgisinin gizli olduğu (Private Equity) durumlarda *Employee Count* tek başına belirleyici olur. En önemlisi, bu verilerin tamamı **Publicly Available (Halka Açık)** kaynaklardan (LinkedIn, Web Sitesi, Google Arama) yasal yollarla toplanabilir; şirketin içeriden veri paylaşmasına gerek yoktur. Bu da modelin "Cold Start" problemini aşmasını sağlar, veriler hemen bulunabilir demek yani.
